

A comparative study of classification techniques in data mining algorithms used for medical diagnosis based on DSS

Ahmed Shihab Ahmed¹, Hussein Ali Salah²

¹Department of Basic Sciences, College of Nursing, University of Baghdad, Baghdad, Iraq

²Department of Computer Systems, Technical Institute-Suwaira, Middle Technical University, Baghdad, Iraq

Article Info

Article history:

Received Sep 16, 2022

Revised Dec 16, 2022

Accepted Jan 9, 2023

Keywords:

C4.5 decision tree algorithm

Classification rules

Clinical DSS

Logistic regression algorithm

Naïve Bayes algorithm

ABSTRACT

A significant amount of data is gathered by the healthcare sector, but it is not appropriately mined and utilized. Finding these hidden links and patterns is frequently underutilized. Our study focuses on this element of medical diagnostics by identifying patterns in the information gathered about kidney illness, liver disease, and chronic pancreatitis (CP) and designing adaptive medical decision support systems (MDSS) to assist doctors. This research compares a variety of data mining (DM) techniques, knowledge extraction tools, and software platforms for usage in a DSS for analysis using the Waikato environment for knowledge analysis (WEKA) mining tool (decision tree (DT)). The objective is to determine the most significant risk factors based on the extraction of the categorization criteria. The datasets used for this work are illustrates how successfully DM and DSS are integrated. In this research, we suggest using the C4.5 DT algorithm, Naïve Bayes (NB) algorithm, and the logistic regression (LR) algorithm to categorize these diseases and evaluate their performance and accuracy rates. It inferred that the C4.5 algorithm accuracy is 0.873% which is better than the other two algorithms in terms of rule generation and accuracy.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Hussein Ali Salah

Department of Computer Systems, Technical Institute-Suwaira, Middle Technical University
Baghdad, Iraq

Email: hussein_tech@mtu.edu.iq

1. INTRODUCTION

Data mining (DM) is the process of detecting patterns in vast amounts of data or uncovering new information from massive amounts of data in terms of patterns or rules. Database technology, analytics, artificial intelligence, pattern recognition, information systems, high-speed computing, and visualization approaches are among the areas involved. Different rules and patterns are mined using DM techniques, such as association rules, sequential patterns, classification trees, and so on. Before it may generate usable information, it must first undergo data preparation. DM main purpose is to extract hidden information from a batch of data. The data gathered is useful in making decisions. Several popular DM technologies are currently being used to successfully find predictive information for a variety of applications. DM results can be presented in a variety of ways, including a list, graphically results, summarized tables, and visualization [1], [2].

The provision of high-quality services at reasonable prices is a major problem for the health sector. A high-quality service entails appropriately diagnosing and treating patients. Poor clinical judgment can have disastrous consequences, this cannot be allowed. Even the most technologically advanced hospitals lack software that uses DM techniques to anticipate sickness. There's a lot of hidden data out there that can be turned into good information. Clinical analysis is recognized in order to be subjective; the doctor providing the analysis determines the outcome. Second, and perhaps most crucially, the quantity of information that must be examined in order to

create a decent forecast is typically enormous and on occasion incomprehensible. Machine learning (ML) is a tool that can be employed in this context to deduce diagnostics autonomously principles based on characterizations of previously patients who have been medicated, allowing doctors to make the diagnosis a more accurate procedure and trustworthy. The decision support system (DSS) created to aid doctors in the diagnosis process are frequently dependent on the basic principle of dataset that is not up to date.

Physicians and hospitals would benefit from a DSS that can understand the correlations between patient's medical background, population illnesses, indications, disease anatomy, personal history, and medical tests. Because of the many different techniques and domains in which decisions are made, the idea of DSS is highly broad. Knowledge-based systems are included in DSSs, which are a sort of computer-based information system. In general, a DSS is a computerized system that assists in decision-making. The subsystems can be used to build a DSS application. However, developing such a system is a difficult undertaking that has yet to be tackled. Many aspects have been discovered, however a major difficulty has been highlighted as a lack of knowledge. To decrease detection time and increase diagnosis accuracy, developing reliable and effective medical MDSS to help the ever-complicated diagnosis decision process has become a more difficult task. Because medical diagnosis is a complicated and ambiguous procedure of reasoning by nature, soft computing techniques like decision tree (DT) classifiers have demonstrated to be effective considerable promise in the development of MDSS for kidney illness, liver disease, and chronic pancreatitis (CP).

Integrated computer-based systems that support decision-making are known as DSS, which assist in identifying, solving, and making decisions by utilizing designs and datasets. They are aimed to aid in semi-structured and unorganized decision-making processes, selection by including both data and models. They supplement rather than replace decision-making. Rather than increasing decision efficiency, the goal of DSS is to increase decision effectiveness [3]. DM enables organizations to make important considerations more quickly and with more assurance. He is convinced that using DM reduces the amount of ambiguity in the decision-making process [4]. Integration of DM can increase DSS performance and allow new kinds of issues to be addressed that have never been addressed previously. They also claim that combining DM and decision support can considerably enhance present methods to issue solving and generate new ones by allowing the merger of expert knowledge with data-driven knowledge [4].

Research by Song *et al.* [5] have develop a fuzzy rule-based categorization system for heart disease using a brand-new data-driven method based on linguistic qualifiers and fuzzy clustering. The proposed system provides a knowledge base that might be utilized to describe the decision-making process. In the experiment, we contrasted the suggested strategy utilizing artificial neural networks, support vector machines (SVM), k-nearest neighbors (K-NN), Naïve Bayes (NB), and random forests (RF) are five well-established ML algorithms for heart disease detection. Heart disease databases from Cleveland, Hungary, and Virginia Long Beach were used. The findings show that, in terms of finding a compromise between precision and interpretability [5].

An up-to-date review of clinical decision support system (CDSS) application in medicine should be provided, covering the various forms, current use cases with proven efficacy, typical problems, and potential drawbacks. They conclude with recommendations that are supported by research for reducing risk in CDSS design, deployment, evaluation, and upkeep [6]. A computerized CDSS that pre-assesses the pertinent data and offers medical providers accurate, helpful suggestions at the point of care has been evaluated by Sutton *et al.* [4] for effectiveness. The hospitals electronic health records were integrated with evidence-based medicine electronic decision support (EBMEDS), a commercial CDSS that creates patient-specific recommendations for a variety of medical illnesses across disciplines. The main result was how frequently medical issues discovered and reported by the CDSS were addressed by a change in procedure. The length of the hospital stays and in-hospital death from all causes were considered secondary outcomes [4]. The primary public medical database was introduced, which also provided a straightforward explanation of the DM process' phases, jobs, and models. They also provided a description of data-mining techniques and their use in real-world scenarios. The purpose of this effort was to assist clinical researchers in developing a clear and intuitive knowledge of the application of data-mining technologies on clinical big-data in order to enhance the creation of research findings that are advantageous to physicians and patients [7].

The goal of this study is to analyze a number of studies on fuzzy logic (FL) and hybrid-based techniques for determining patient heart disease risk. The analysis provides a collection of studies from 2010 together with the power, operating system, accuracy rate, and other specifications used in FL and hybrid-based approaches for the diagnosis of heart illness [8]. The primary goal of this research is to examine categorization methods utilizing the Waikato environment for knowledge analysis (WEKA) ML platform. Additionally, a sizable dataset was utilized. The field of protein structure prediction produced this dataset. It has already been divided using the ten-fold cross-validation method into training and test sets. 9 different techniques have been tested in this trial. It became clear as a result that testing more than one classifier from the tree family in the same experiment is not appropriate. However, utilizing the NB classifier with the

attribute selection filter's default properties takes a lot of time. Finally, it is important to prioritize changing the attribute selection parameters for more accurate results [9].

In-hospital mortality and internal medicine (transfer to a hospital) have been the results (direct hospital admission or transfer). We created four ML-based models: lasso regression, RF, gradient-boosted DT, and deep neural network using commonly accessible triage data as predictors (such as demographic information and vital signs in the training set (70% random sample)). We assessed the predictive performance of the models using statistics, prospective prediction values, and decision curves for the test set (the remaining 30% of the data). These ML models were developed for each result using the common triage categorization data and evaluated to the model [10]. The patients with heart disease have been predicted using a variety of DM approaches. However, using DM techniques did not eliminate the data's ambiguity. Fuzziness has been added to the measured data in an effort to reduce uncertainty. To eliminate ambiguity, a membership function was created and added to the measured value. Additionally, an effort was made to categorize the patients using the data gathered from the medical community. The minimum distance K-NN classifier was used to divide the data into different categories. In comparison to other classifiers of parametric approaches, it was discovered that the fuzzy K-NN classifier works well [11].

In this study, a variety of classification algorithms based on factors like age, gender, blood pressure, cholesterol, and pulse rate are used to evaluate each person's risk level. DM techniques such as NB, K-NN, DT algorithm, and neural network are used to categorize the patient risk level. The risk level can be predicted with a high degree of accuracy when more criteria are employed [12]. The massive amount of data generated by the healthcare industry, ML algorithms significantly contribute to the disease prediction. The leading cause of death in India is heart disease. According to WHO, timely steps can anticipate and prevent stroke. By using ML approaches like DT and NB as well as risk variables, the study in this paper can assist predict cardiovascular disease with greater accuracy. The heart failure dataset, which has 13 attributes, is the dataset that we took into consideration. Pre-processing of the gathered data is necessary before examining the performance of approaches. Then feature selection and reduction should come next [13]. The medical industry has a lot of promise for using DM to find patterns that are hidden in medical datasets. This makes it possible to abstract knowledge for predicting heart disease using a variety of mining techniques. A survey of different single DM approaches and hybrid mining techniques is conducted to determine the most effective method for achieving high accuracy in heart disease prediction. Here, the potential of a variety of classifying strategies, including NB, SVM, DT, K-NN, and even a hybrid classifier approach, was assessed. Analysis of several methods showed that classification-based techniques outperform earlier strategies in terms of accuracy. DM, classification, disease diagnosis, forecasting, and accuracy are some related terms [14].

This article's purpose is to determine the most important risk factors for managing the kidney, liver, and CP disorders issue and to provide diagnoses, therapies, and invalidations for each level among them. DSS provides tools for measuring infection levels, locating solutions, and providing short instructions or tasks. Additionally, it demonstrates appropriate step and care taken to prevent deaths, aids medical staff members working in hospitals, and concentrates on finding information sources. The primary goal of this paper is to improve the DSS based on DT C4.5 inference system (DTCIS), NB algorithm, and logistic regression (LR) algorithm to diagnose these illnesses and compare their effectiveness and correction rates, an algorithm was developed to gracefully manage requests for medical care in order to reduce the symptoms of CP, kidney disease, and liver disease, as well as to control medical care flexibly in order to offer an optimal way of treatment.

The organization of this manuscript is as follows. Section 2 discussed the relevant related papers discussed the use of classification technologies in the healthcare DSS field are surveyed. In section 3 discuss the importance of medical diagnosis detection using ML. In section 4 we analysis the results. Section 5 explained all the information that related with the experimental data of diseases. In section 6 depicts the DM algorithms. Section 7 describes the performance evaluation. Section 8 illustrates the comparison of NB, C4.5 and LR algorithms. Section 9 describes the conclusions.

2. RELATED WORK

Several studies focusing on medical diagnostics have been published so far. Using the UCI ML repository's dataset, these studies used several solutions to the issue and produced great efficiency in classification of 77 percent or more [15]. Some instances are as follows: with a logistic-regression-derived discriminant function, experimental results indicated a proper classification accuracy of roughly 77%. To detect cardiac illness [16]. Bahani *et al.* [17] have employed fuzzy support vector clustering. The experimental findings were achieved using a well-known benchmark of heart illness, and the algorithm used a measurement created by a kernel to allocate every piece of information. Support for ischemic heart disease (IHD). Vector machines are good forecaster and detectors with a high degree of accuracy. Nonlinear proximal SVM (PSVM) are used in this tree-based classifier. Kaur and Khehra [18] have using a principal

component analysis, an expert system was created to diagnose diabetes. A cascade learning system was also developed to identify diabetes. Developed a fuzzy-based controller to regulate blood glucose levels using expert knowledge. To obtain variations from a dataset of self-monitoring blood sugar levels, devised a stochastic model [19]. Salah and Ahmed [20] have design a DSS application to assist specialists (doctors) in making challenging decisions.

The DSS is based on specialists' experience and a DM extraction strategy to assist the hospital handle the (COVID-19) viral pandemic gracefully and, more broadly, to define the type of disease and give a suitable protocol health indicator on the diagnosis. To begin, it is necessary to identify the 3 early COVID-19 pandemic diagnosis (fever, weariness, dry cough, and breathing problems) that are used to establish whether or not a person is infected with the virus. Second, employing two age factors and primary healthcare status variables like diabetes, heart problems, or ischemia, this approach separates infected persons into several categories depending on their immune response risk (very high degree, high degree, mild degree, and normal). Where these folks are assessed and expected to follow the rules of their class. The major goal of this work is to improve a DSS based on DTCIS to elegantly monitor requests in medical care to decelerate COVID-19 virus symptoms and regulate a pandemic flare-up to reduce its impact on medical care. utilizing a good treatment protocol to deliver a proper therapy. Ahmed *et al.* [21] have design a first aid decision support system (FADSS) and built it to provide access to practical instances that pose a risk to the general public, as well as sophisticated conditions for assessing research abilities and providing for critical medical care via a graphical user interface. The design of FADSS's first-aid therapy is based on common first-aid scenarios. We provided an approach for managing first-aid therapy by modeling a framework (FADSS) that helps individuals access data about first-aid situations that are commonly accessible as a service. The FADSS service employs a set of fifteen critical scenarios that may occur in individual's lives. A therapy decision is recommended by a technique for a new kind of disaster. FADSS uses computational models, DT, and DM (C4.5 algorithm) to test information in real-time in order to develop a decision-making system. When the case is really critical, the system sends out automatic alerts via text messages and email reports. The primary goal of this research is to develop an effective tool that assists individuals and junior staff at first-aid centers in locating relevant information resources.

Although detection for emphysema using low-dose computed tomography reduces lung cancer mortality, it also has the potential to cause harm. Patients should be informed as to the advantages and risks of detection for emphysema before making a choice, according to current guidelines. To compare the impact of a decision-making assistance for patients on detection for emphysema decision-making outcomes vs professional requirements material (EDU) among smokers. When compared to EDU, a patent ductus arteriosus (PDA) sent to people searching help from smoking stop lines enhanced the quality of detection for emphysema decisions. These advancements were in line with professional society advices for tobacco making educated decisions about detection for emphysema [22].

The patient decision aid had no effect on whether or not the patient intended to be screened or whether or not the screening was completed. Motivations to get screened or screening results were not affected by the PDA. But there is a critical necessity enhance the quality of life of these talks, the patient decision aid was designed to supplement but not as a replacement for a medical discussion practitioner. 44 the PDA may reach a huge a large number of possibilities qualified tobacco in the US if it was distributed through tobacco stop lines. Given the varied funding for cigarette stop lines, treating the function of lines for quitting smoking in spreading patient decision aid support for detection for emphysema is critical to ensure that the intervention is widely disseminated and has a greater impact.

Medical evidence can be filtered derived from research, integrated using computerized patient history, and advices targeted to particular patients generated using sophisticated evidence-based information resources. The goal of this study was to see how effective an electronic CDSS is in reappraising indication and providing health care practitioners provides patient-specific, relevant data suggestions at this time of treatment. However, the modification despite the fact that an investment arbitration clinical decision support therapy was only statistically successful in improving providing activity and service quality, this pragmatic randomized controlled trials (RCT) found that it was marginally useful in changing prescribing behaviour and service quality. Impacts on improving adherence to suggestions were small, influencing only about 4 out of every 100 patients. An electronic health record with a standardized conceptual and layout system that can be easily managed. complicated information on disease, 46-48 and the capacity to link data to a CDSS that is either supplier or accessible are the minimal criteria for adopting CDSSs. 49,50 these essential elements are increasingly prevalent preconditions for CDSS use in hospitals, and they can help to promote and sustain CDSS implementation in health authorities. A DSS is additional feature of a health-monitoring system [23].

A DSS is a platform that collects data from many resources and current events it in creative visual representations such as graphs, geographical analyses, and maps, as well as reports from the health-care industry and other sources of information. A DSS transforms health data into actionable information more accessible, intelligible, and, as a result, decision-makers are more motivated to use it. Measure evaluation's

prototype DSS is capable of integrating many different data sources and is a strong but user-friendly tool for data analysis. We must ensure that data sources are compatible as they are strengthened and additional become accessible. The MFL is the most essential data source since it allows data sources to be linked. As a result, it necessitates extra attention. The DSS provides decision makers with comprehensive data for monitoring programs and avoiding and controlling outbreaks.

There have now been a few investigations that have focused on clinical determination. Using the dataset acquired from the user computer interface AI archive, these investigations used a variety of approaches to address the problem and achieved high grouping correctness's of 77% or higher [24]. Here are a few examples of models: fuzzy support vector clustering was used to differentiate coronary artery disease. To relegate each bit of information, this method used a portion initiated measurement, and exploratory the outcomes were acquired through using a well-known standard of cardiovascular sickness. Nonlinear proximal assistance vector machines are used in this tree-based classifier [25]. Based on synthesis analysis, Aljohani *et al.* [26] have established a professional framework to analyse the diabetic sickness. To analyse diabetes, he created a course learning architecture. Also enhanced a fluffy-based regulator that combines expert information to control the sugar levels in the blood levels.

The adequacy of CDSS has been investigated in a number of different studies. The framework and analytic initiatives advise doctors on potentially dangerous pharmaceutical combinations. These projects can help to limit concerns and blunders, avoid misunderstandings, and improve the doctors' analyses. Early warning in the event of harm may have an impact on the type of care provided and the expense involved [27]. According to a study conducted in England, executing PC-based rules can result in an improvement in health outcomes, and the unresolved questions raised by clinicians throughout the clinical experience will provide an opportunity to use the CDSS [28]. They compiled the four factors associated with implementing successful CDSS from several studies. The elements were: i) automating alarms and updates, ii) providing proposals at a specific location and time, iii) providing substantial suggestions, and iv) automating the complete process. These elements have an impact on the process of crisis care and treatment. To demonstrate how quickly the DSS communicated the issue, a contextual analysis approach was used. Data sharing is a critical component of properly implementing a CDSS [29].

The findings reveal that doctors accept losing control of their work and losing special aptitudes and information by following the advice of CDSS, where any non-expert can gain access to clinical material specified by the doctor. As a result, skilled self-sufficiency plays an important role in doctors' decision to use a CDSS. Furthermore, this investigation improves; i) structure, which encourages the chief to assign a domain to doctors in order to facilitate effective information sharing and the implementation of intuitive CDSS, and ii) the nature of administrations provided to patients through the use of appropriate clinical information technology (IT) frameworks in clinics. Antoniadi *et al.* [30] have looked into the most significant CDSS issues. These include computerization of the entire CDSS, integration into clinical work processes, framework extensibility and viability, ideal advising, cost-benefit analysis, and the requirement for structures that allow for the reuse and able to share of CDSS administrations and components.

3. MEDICAL DIAGNOSIS DETECTION USING MACHINE LEARNING

3.1. Machine learning techniques

ML techniques are now being utilized to identify and defend outliers, especially at the detection stage. Algorithms such as the SVM, K-NN, neural network, DT, NB, and others are now in use. ML is a collection of algorithms that convert data into actionable information. When it complements rather than replaces a topic master's unique knowledge, it works well. As the name implies, a predictive model is used to forecast one value based on the dataset's other values. The learning algorithm seeks to deduce and simulate the relationship between the goal and other characteristics. The processing of a training predictive model is referred to as supervised learning or classification [31]. DT, NB, LR, and RF are examples of supervised learning approaches (RF). In this study, we develop four ML models using the LR, NB, and RF, DT, algorithms, which we then analyze to discover the best model. The C4.5 DT algorithm is a DT that is used to make decisions.

This method is a better version of his prior C4.5 (j48) algorithm, which was better than his iterative measures of depression 3 technique (LR). The C4.5 algorithm has the advantage of being opinionated when it comes to trimming and making a lot of decisions automatically with very good defaults. The C4.5 algorithms make use of the information entropy concept. The approach requires a set of input and output training pairs, with the appropriate class as the output. The result is displayed as a tree, making it human-readable. It has a variety of characteristics, including [31]: i) the C4.5 approach is capable of detecting noise and missing data, ii) the C4.5 approach is capable of detecting noise and missing data. The huge DT can be thought of as a set of straightforward rules, iii) the C4.5 classifiers can forecast which attributes are relevant and which are not when it comes to categorization, and iv) overfitting and pruning errors were no longer a problem.

3.2. Feature selection

To improve classification performance and save memory, feature selection is a procedure for choosing a subset of significant characteristics from a greater number of features and reducing the amount of irrelevant redundant features in a dataset [32]. Feature extraction aids in data interpretation, reduction of the curse of dimensionality, decrease of processing requirements, enhancement of learning accuracy, and distinction of which features may be essential to a particular situation [33]. There are several supervised feature selection strategies, which can be split into wrapper, filter, and embedding models. One of the most commonly used filter model approaches in the training dataset of each attribute is examined by assessing the utility of an attribute fuzzy with regard to the class in feature selection (Figure 1). The higher a characteristic's entropy, the more information rich it is [33].

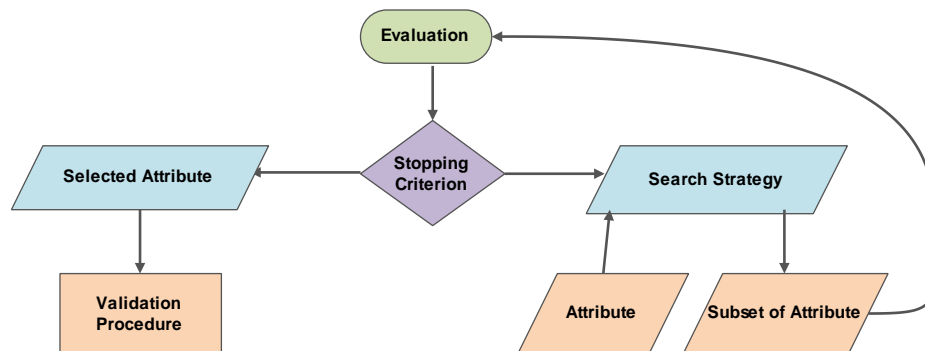


Figure 1. Flowchart of a feature subset selection

3.3. Feature weighting

Feature weighing is a viable approach to keeping or eliminating a feature. The more important traits are given greater weight, while the less important features are given less weight. Large-weighted features play an essential role in the model's construction, resulting in improved accuracy. The domain's knowledge of the relative relevance of features is frequently used to determine these weights. Alternatively, it could be chosen for you automatically [34].

4. ANALYSIS AND RESULTS

4.1. The dataset

The purpose of this project is to create and evaluate a CDSS for the management of patients with kidney disease, liver disease, and CP. Kidney illness is the greatest cause of death worldwide, according to one survey. Almost 830,000 individuals die in the United States alone, at a cost of about 393.5 billion dollars. The human kidney is positioned a cross on both sections of the lumbar spine on the posterior abdominal wall. The kidney's key tasks include metabolic management, waste and toxin excretion, blood pressure regulation, and fluid balance maintenance. The kidney filters all of the blood in the body 20 times every hour. Whenever kidney failure is reduced, the waste of the body is unable to be digested, resulting in back discomfort, hypertension, kidney disorders, high blood pressure, urethral inflammation, lethargy, sleeplessness, vertigo, hair loss, and eyesight blur, poor reaction time, sadness, fear, mental problems, and other complications. A damaged kidney will also generate and secrete erythropoietin. Patients will get anemia if their red blood cell production is insufficient. Kidney failure patients may present bone fractures because the kidney helps regulate the calcium and phosphate [35]. In the United Kingdom, 700,000 people have symptomatic kidney failure. Although there is a tendency toward bettering the prognosis of kidney failing patients, mortality is still significant, especially in the first year. This is despite breakthroughs in renal failure treatment, which include anything from effective medication management to invasive procedures like cardiac resynchronization therapy (CRT). As a result, continuous research into ways to identify patients who are more likely to die early is justified. The predictive value of patients with poor liver function tests (LFTs) with newly diagnosed renal failure with lower elimination proportion is one such route (KFREF).

LFTs were abnormal in patients with right renal failure and pericardial regurgitation as a result of liver obstruction, primarily alkaline phosphatase (ALP) and bilirubin. Similarly, ischemic hepatitis caused by cardiogenic shock results in aberrant aminotransferases. Low albumin levels, which indicate insufficient nutrition, are common in patients with cardiac cachexia. To avoid acute consequences and reduce the risk of long-term complications, hepatitis, a liver illness, necessitates ongoing medical care and patient self-management education.

Anorexia (lack of appetite) and an increase in alkaline phosphate levels are to blame for this. Hepatitis is one of the diseases that can be categorized [36].

Fibrous tissue is increasingly replacing endo-and exocrine pancreatic cells as acinar structures are depleted characterize CP. Pancreatic exocrine insufficiency frequently occurs before endocrine insufficiency in the course of CP. Digestive enzyme secretion and, as a result, activity decreases over time, resulting in maldigestion. Weight loss is common in CP patients related to a decrease in food consumption (due to pain and/or long-term alcohol consumption) and proteins maldigestion, complex acids, and carbs. Malnutrition is therefore widespread in CP patients, and its severity is one of the most important indicators in predicting complications and the disease's fate. In a recent study, we discovered that patients with advanced CP have significantly worse nutritional indicators. Protein calorie deficiency has been linked to lower serum amino acid concentrations [37].

4.2. The WEKA tool

WEKA is a collection of DM ML algorithms. The algorithms can be used with a collection of data or used straight from Java code. WEKA includes data pre-processing, classification, regression, data collecting, association rules, and visualization tools. It's also ideal for creating new ML strategies [38]. The ability to read files from a number of database formats is one of WEKA's strongest features [39]. A visualization support is the software's flaw. It's worth emphasizing that the software serves as a visual representation of data, results, and procedures. The assistance available is somewhat restricted. WEKA can communicate with the R statistical package. In order to improve not only the statistical analysis operations, but also the visualization of statistical analyses and findings [40]. WEKA is a free open source application distributed under the terms of the GNU general public license. WEKA was originally developed in C, but it has since been fully rewritten in Java and is incompatible with nearly every computing platform. WEKA is simple to set up and use, featuring a graphical interface that allows for quick setup. WEKA is applicable to a wide range of DM techniques and industries. Users can use the application to find hidden information databases, files with user-friendly interfaces, and visualizations [41], [42].

5. EXPERIMENTAL DATA

Kidney illness, liver disorders, and CP were the three medical datasets we used. All of these datasets were taken from the ML datasets archive. The goal is to categorize illnesses and assess features extraction measure techniques like NB, C4.5, and LR. This experiment uses the kidney illness dataset of 340 individuals, which has 29 sequentially valued and important features, as indicated in Table 1. The liver diseases dataset contains 23 attributes and 220 instances, as shown in Tables 2 and 3. Table 4 shows the CP dataset comprising 680 individuals with 27 variables [43]. Table 3 describe the elements of liver function and their normal values used in this study are prescribed to compare them with the values achieved by the liver patient. Table 4 describe the elements of CP function and their normal values used in this study are prescribed to compare them with the values achieved by the liver patient.

5.1. Attributes selection measures

ML and DM use a variety of measures to develop and evaluate models. The DT C4.5 methods, the NB algorithm, and the LR algorithm have all been developed and tested on our experimental datasets. The confusion matrix created by these algorithms can be used to assess their correctness precision, recall, F-measure, and receiver operating characteristic (ROC) space were the four performance measurements used [44]. The four measures are calculated using a distinct confusion matrix (also known as a contingency table). The classification results are represented as a matrix in the confusion matrix. It includes information on a categorization system's actual and expected classifications. The data set classified as true when they were exactly true (i.e., TP) and the number of variables categorized as false when they were clearly false (i.e., FP) (i.e., TN). The number of samples that were erroneously classified is represented by the other two cells. Furthermore, the cell indicating the number of tests categorized as false while they were exactly true (i.e., FN) and the cell representing the number of results generated during the testing as true when they were clearly false (i.e., TF) (i.e., FP). Following the generation of the confusion matrices, the precision, recall, and F-measure may be easily calculated.

$$Recall = TP / (TP + FN) \quad (1)$$

$$Precision = TP / (TP + FP) \quad (2)$$

$$F_measure = (2 \times TP) / (2 \times TP + FP + FN) \quad (3)$$

Precision, in less formal terms, assesses the proportion of patients who are actually sick (i.e. true positive) among the patients who were proclaimed disease. Recall, on the other hand, measures the percentage of actually sick who were detected; and F-measure, on the other hand, balances precision and recall. False positive rate (FPR) and true positive rate (TPR) are the x and y axes, respectively, in a ROC space, which illustrates true positive and false positive tradeoffs in relation to each other.

$$TPR = TP / (TP + FN) \quad (4)$$

$$FPR = FP / (FP + TN) \quad (5)$$

Table 1. Demonstration of the attributes in the kidney disease dataset [35]

No	Attribute name	Description
1	Blood urea nitrogen (BUN)	Reference/units 5–25/ mg/dL
2	Creatinine (CRE)	0.3–1.4 mg/dL
3	Uric acid (UA)	2.5–7.0 mg/dL
4	Albumin-globulin in ratio (A/G ratio)	1.0–1.8
5	Creatinine clearance/24 hrs urine (CC)	M: 71–135 F: 78–116 mL/min
6	Renin (Penin)	0.15–3.95 pg/mL/hr
7	Creatinine urine (Creatinine urine)	60–250 mg/dL
8	Sodium (Na)	135–145 meq/L
9	Potassium (K)	3.4–4.5 meq/L
10	Calcium (Ca)	8.4–10.6 mg/dL
11	Phosphorus (IP)	2.1–4.7 mg/dL
12	Alkaline phosphatase (ALP)	27–110 U/L
13	Hemoglobin (Hb)	M: 14–18 F: 12–16 g/dL
14	Red blood cell (RBC)	M: 450–600 F: 400–550 mil/mm ³
15	White blood cell (WBC)	5000–10000 mm ³
16	Hematocrit (Hct)	M: 40–55 F: 37–50%
17	Platelets (PLT)	15–40.0 103/uL
18	Mean corpuscular volume (MCV)	83–100 u3
19	Mean corpuscular hemoglobin (MCH)	27–32.5 uug
20	Mean corpuscular hemoglobin concentration (MCHC)	32–36%
21	Reticulocyte (Reticulocyte)	0.5–2.0%
22	Malaria (Malaria)	(–)
23	Erythrocyte sedimentation rate (ESR)	M: 1–15 F: 1–20 mm/hr
24	Neutrophils (Neutrophils)	50–70%
25	Lymphocytes (Lymphocytes)	20–40%
26	Bleeding times (BT)	0–3 minute
27	Blood pressure (BP)	mm/Hg
28	Coagulation times (CT)	2–6 minute
29	Class	1 means (kidney test positive) while 0 means (kidney test negative).

Table 2. Demonstration of the attributes in the liver diseases dataset [36]

No	Attribute name (characteristic)	Normal group (n=124)	Abnormal group (n=95)	p-value
Demographic and clinical				
1	Age	69.9 ± 12.3	68.6 ± 14.3	0.46
2	Gender (M)	85 (68.5%)	66 (69.5%)	NS
3	Ethnicity (white)	99 (79.8%)	74 (77.8%)	NS
4	NYHA 2	71 (57.3%)	50 (52.6%)	NS
5	NYHA3/4	33 (26.7%)	31 (32.6%)	NS
6	EF	35.6 ± 11.5	36.7 ± 13.8	0.5
Medical history				
7	CAD	53 (42.7%)	38 (40%)	NS
8	Hypertension	70 (56.5%)	51 (53.7%)	NS
9	Diabetes mellitus	32 (25.8%)	21 (22.1%)	NS
10	Atrial fibrillation	30 (24.2%)	34 (35%)	NS
11	Paced	12 (9.7%)	9 (9.5%)	NS
12	Smoker	44 (35.5%)	38 (40%)	NS
Laboratory testing				
13	Bilirubin total	11.5 ± 6.5	20.9 ± 17	<0.001
14	ALT	21 ± 8	108.8 ± 327.6	0.004
15	ALP	71 ± 19.6	169.3 ± 169	<0.001
16	Albumin	41 ± 6	37 ± 6.8	<0.001
17	Sodium mmol/L	137.2 ± 9.5	140 ± 3.8	0.06
18	Creatinine class (μmol/l)	116.9 ± 83.9	136.8 ± 93.8	0.1
19	Class			0.1

Table 3. Liver diseases dataset liver function tests with their normal values as used in the study [36]

No	Liver function tests	Normal values
1	ALT	10 IU/L-40 IU/L
2	ALP	40 IU/L-130 IU/L
3	Albumin	35 g/dl-50 g/dl
4	Bilirubin	5 mg/dl-25 mg/dl
ALT: alanine aminotransferase; ALP: alkaline phosphatase		

Table 4. Specified analytical values in chronic pancreatitis patients and healthy controls [37]

No		Control	Chronic pancreatitis	Statistical significance
1	No. of patients	21	35	
2	Age (yr)	34 ± 13	50 ± 10	< 0.01
3	Body weight (kg)	76 ± 4	64 ± 10	< 0.01
4	Body mass index (kg/m ²)	25 ± 2.7	22 ± 3.4	< 0.01
5	Serum total protein (g/L)	76 ± 4	72 ± 8	< 0.05
6	Serum albumin (g/L)	46 ± 3.6	42 ± 8	< 0.05
7	Serum triacylglycerol (mg/dL)	107 ± 30	119 ± 43	NS
8	Serum cholesterol (mg/dL)	195 ± 26	175 ± 44	< 0.05
9	Glucose (mmol/L)	5 ± 0.64	7.7 ± 2.8	< 0.01
10	Insulin (μU/mL)	12 ± 5.2	9.1 ± 4.4	< 0.05
11	Homeostasis model assessment	2.7 ± 1.3	2.6 ± 1.6	NS
12	Leptin (ng/mL)	6.6 ± 2.9	3.9 ± 2.5	< 0.01
13	Hemoglobin (mg/dL)	15 ± 2.9	13 ± 1.8	< 0.01
14	Blood urea nitrogen (mg/dL)	11 ± 2.8	14 ± 6	NS
15	Creatinine (mg/dL)	0.9 ± 0.13	0.89 ± 0.22	NS
16	Serum amylase (U/L)	54 ± 31	65 ± 35	NS
17	Urine amylase (U/L)	174 ± 66	344 ± 336	< 0.05
18	Serum lipase (U/L)	28 ± 16	69 ± 75	< 0.05
19	Serum alanine	21 ± 9	38 ± 38	< 0.05
20	aminotransferase (U/L)			
21	Serum aspartate	19 ± 5	41 ± 45	< 0.05
22	aminotransferase (U/L)			
23	Serum alkaline	80 ± 28	104 ± 65	NS
24	phosphatase (U/L)			
25	Serum-glutamyl	26 ± 14	112 ± 152	< 0.05
26	transpeptidase (U/L)			
27	Class	0.1		

6. DATA MINING ALGORITHMS

Different algorithms are used in DM to transform raw data into knowledge that can be applied. A classification method is used in vocations when it is necessary to forecast one value using data from other values in the dataset, as the name suggests. The learning algorithm tries to infer and simulate how the aim and other features relate to one another. Supervised learning or categorization is the process of processing a training predictive model. DT, NB, neural networks, supervised learning techniques include SVMs and RF, for instance [45].

Four models were created in this study in order to get the results. After contrasting NB, RF, DT, and LR, the best model was chosen. A powerful method for data exploration, predictive modeling, and analysis that is widely used is RF. Individual DT that are created from a number of separately trained DT can deliver results (RF). A classification model uses learning techniques to generate a group of classifiers, and it then uses a grading system for predictions to classify new data. The output is done using individual trees in the RF approach, which comprises of multiple DT [46]. It has a number of features, including: i) offers excellent and efficient services for missing data and methods for dealing with it, and ii) because over processing is an issue in some DT, this strategy is the best answer [46].

6.1. Naïve Bayes

This procedure is the foundation of bayes theory and is employed when the number of input dimensions is large. The output of a bayesian classifier can be computed from the input. In addition, you can upload fresh data at any time during the game and gain points for the best probability classifier. When the class variable is given, the presence (or absence) of a feature assigned to a class, according to a NB classifier, is unrelated to the presence (or absence) of any other characteristic, as demonstrated in Tables 5-7 [47].

Table 5. Matrix of perplexity of NB algorithm-kidney disease dataset

TP rate	FP rate	Precision	Recall	F-measure	ROC area	Class
0.786	0.381	0.76	0.786	0.773	0.78	No
0.819	0.414	0.842	0.819	0.83	0.819	Yes

Table 6. Matrix of perplexity of NB algorithm-liver disease dataset

TP rate	FP rate	Precision	Recall	F-measure	ROC area	Class
0.786	0.381	0.76	0.786	0.773	0.78	No
0.819	0.414	0.842	0.819	0.83	0.819	Yes

Table 7. Matrix of perplexity of NB algorithm-chronic pancreatitis disease dataset

TP rate	FP rate	Precision	Recall	F-measure	ROC area	Class
0.686	0.181	0.76	0.683	0.743	0.78	Yes
0.919	0.514	0.832	0.912	0.931	0.789	No

6.2. Logistic regression

Is a predictive analysis technique that works in the same way as other regression analyses. One of the responsibilities of LR is also to describe the data. The ensuing link between classes and attributes is explained and shown in (Tables 8-10) [48].

Table 8. Matrix of perplexity of LR algorithm-kidney disease dataset

TP rate	FP rate	Precision	Recall	F-measure	ROC area	Class
0.681	0.486	0.64	0.682	0.671	0.78	No
0.713	0.517	0.748	0.713	0.76	0.78	Yes

Table 9. Matrix of perplexity of LR algorithm-kiver disease dataset

TP rate	FP rate	Precision	Recall	F-measure	ROC area	Class
0.882	0.721	0.91	0.93	0.973	0.784	Yes
0.412	0.04	0.742	0.719	0.63	0.784	No

Table 10. Matrix of perplexity of LR algorithm-chronic pancreatitis disease dataset

TP rate	FP rate	Precision	Recall	F-measure	ROC area	Class
0.688	0.191	0.723	0.589	0.673	0.842	Yes
0.919	0.482	0.848	0.878	0.826	0.842	No

6.3. Decision tree C4.5 algorithm

DM is one of the most essential techniques, DT are used in measurements, calculations, and ML. Using a DT, one can move from a particular idea (represented by branches) to conclusions regarding the utility and worth of an object (as an insight model) (represented as leaves). Class labels refer to the leaves, and conjunctions of climax refer to the branches that symbolize the class labels. Regression trees are DT in which the objective variable can take on persistent features (often true numbers). Because of its comprehensibility and clarity, DT are one of the most commonly used DM tools [48]. C4.5 chooses one data characteristic at every node of the tree that splits its set of samples most efficiently into subgroups enhanced in one or the other category [47]. The normalized information gain (difference in entropy) that results from selecting a property for data splitting is its criterion. To make the decision, the attribute with the highest normalized information gain is picked. LR was improved by C4.5 [48].

- Taking care of both continuous and discrete properties. Creates a threshold and then divides the list into those whose attribute value is more than or equal to the threshold and those whose attribute value is less than or equal to it
- Working having input parameters that are missing in training data
- Dealing with qualities that have different costs
- Pruning trees after they've been made. C4.5 goes back through the tree after it's been created and replaces branches that don't help with leaf nodes. When the C4.5 methods is used to the three medical datasets [48]. The results are shown in Tables 11-13.

Table 11. Matrix of perplexity of DT C4.5 algorithm-kidney disease dataset

TP rate	FP rate	Precision	Recall	F-measure	ROC area	Class
0.786	0.281	0.786	0.786	0.786	0.832	No
0.796	0.293	0.796	0.796	0.796	0.832	Yes

Table 12. Matrix of perplexity of DT C4.5 algorithm-liver disease dataset

TP rate	FP rate	Precision	Recall	F-measure	ROC area	Class
0.943	0.781	0.867	0.943	0.974	0.632	Yes
0.314	0.094	0.948	0.816	0.373	0.632	No

Table 13. Matrix of perplexity of DT C4.5 algorithm-chronic pancreatitis disease dataset

TP rate	FP rate	Precision	Recall	F-measure	ROC area	Class
0.632	0.221	0.896	0.983	0.773	0.823	Yes
0.914	0.514	0.757	0.843	0.93	0.823	No

6.4. Classification rules

Considerable standards are obtained, that are important for understanding the experimental dataset's data pattern and actions [49]. Using the C4.5 DT algorithm, the following pattern was discovered [50]. The following are some of the rules retrieved from the kidney disease dataset:

- Kidney disease (absence): creatinine=0.3–1.4 mg/dL, alkaline phosphatase=27–110 U/L
- Kidney disease (presence): mean corpuscular volume=83–100 u3, mean corpuscular hemoglobin concentration= 32–36%, reticulocyte =0.5–2.0%, neutrophils =50–70%
- Kidney disease (absence): natrium =135–145 meq/L, potassium =3.4–4.5 meq/L, calcium=8.4–10.6 mg/dL, phosphorus=2.1–4.7 mg/dL

The rules for liver disease are some of the datasets that have been obtained as following:

- Age 68.6 ± 14.3 .
- NYHA 2=50 (52.6%), NYHA3/4=31 (32.6%), EF=36.7 \pm 13.8.
- Hypertension=51 (53.7%), diabetes mellitus=21 (22.1%), atrial fibrillation=34 (35%)
- Bilirubin total=20.9 \pm 17, ALT=108.8 \pm 327.6, ALP=169.3 \pm 169

There are certain categorization criteria for CP disease datasets are as follows:

- Age (yr)=50 \pm 10, body weight (kg)=64 \pm 10, serum total protein (g/L)=72 \pm 8
- Serum albumin=42 \pm 8, serum triacylglycerol=119 \pm 43, serum cholesterol (mg/dL)=175 \pm 44
- Glucose=7.7 \pm 2.8, insulin=9.1 \pm 4.4, leptin=3.9 \pm 2.5, hemoglobin=13 \pm 1.8
- Blood urea nitrogen=14 \pm 6, creatinine=0.89 \pm 0.22, serum amylase=65 \pm 35
- Urine amylase=344 \pm 336, serum lipase=69 \pm 75, serum alanine=38 \pm 38, serum alkaline=104 \pm 65

7. PERFORMANCE EVALUATION

The many types of classification errors have an effect on its strength in one way or another. We must pay close attention to the expenses associated with errors. As a result, we evaluated the data we obtained using a set of standard indicators: positive either accuracy or value (Pr). It is the ratio of correctly classified attack flows (TP) to properly classified flows for all attributes (TP+FP). Sensitivity or recall (Rc) it measures the ratio of correctly categorized attribute flows (TP) to all properly classified attribute flows (TP+FN). The detection rate is the frequency of actual events that can be predicted to occur. The accuracy, precision, recall, and F1 were calculated using (6) and (7) [44].

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

$$F - \text{score} = \frac{2 \times P \times R}{P+R} \quad (7)$$

NB, RF, DT, and LR are the four techniques for DM that are most widely used. The performance test results of our evaluation measures for these four methods are shown in Table 14. These results depend on the confusion matrices Table 14, in addition to the performance measurement (1)-(4). With an accuracy rate of 87.3% and 87.3%, respectively, and a potential success (precision) of 89% for them, DT and RF classifiers are superior to the others among the three classification algorithms for handling numerical data in particular that were assessed. Since this experimental approach is more optimal, the F1 scor to DT and RF are 88.19% and 87.47%, respectively.

Table 14. Prediction accuracy table

Model	Accuracy	Recall	Precision	F1 score	Time consumer
Naïve bayes algorithm	0.642	0.6693	0.4114	0.5438	18 S
Decision tree C4.5 algorithm	0.873	0.8747	0.8900	0.8819	35 S
Logistic regression algorithm	0.741	0.7695	0.79812	0.7761	1.51 m

8. COMPARISON OF NB, C4.5 AND LR ALGORITHM

When it comes to category descriptions, aggressive, divide-and-conquer tactics are used, algorithm designers have had a lot of success. For this survey, DT learners such as NB, C4.5, and LR were chosen since they're reasonably quick and create classifiers that compete. Evaluating these three methods' ambiguity matrix (Table 15), we discovered that while LR outperforms the NB algorithm on Indices of feature extraction, C4.5 outperforms both in terms of accuracy and time complexity. The run time complexity of LR is satisfactory when compared to C4.5 [50].

Table 15. Comparison accuracy table

No.	Name of algorithm	Accuracy (%)
1	Naïve bayes algorithm	0.642
2	Decision tree C4.5 algorithm	0.873
3	Logistic regression algorithm	0.741

9. CONCLUSION

One of the really efficient categorizations algorithms is the decision-tree algorithm. The algorithm's effectiveness and rate of rectification will be determined by the data. Each model's confusion matrix was computed using 8-fold cross validation, and the performance was evaluated using precision, recall, F measure, and ROC space. Bagging algorithms, particularly C4.5, performed best among the examined approaches, as expected. The findings presented here make practical application more accessible, paving the way for significant progress in the treatment of kidney, liver, and CP. The survey looks at the number of steps involved in data processing and the challenge of running data for the DT algorithms LR, C4.5, and NB. It may be inferred that the C4.5 algorithm outperforms the other two algorithms in terms of rule generation and accuracy. This shown that the NB algorithm outperforms the LR and NB algorithms in terms of induction and rule generalization. The results are then saved in the decision support repository. Since then, the knowledge base has narrowed to a specific group of disorders. The approach has been confirmed through a case study, and the extent of modeled medical knowledge can be expanded. Furthermore, interactions between the patient's various drugs should be explored in order to improve decision support.

ACKNOWLEDGEMENTS

Authors would like to thank Baghdad University, College of Nursing, Department of Basic Sciences and Middle Technical University, Technical Institute-Suwaira, Department of Computer Systems for the support given during this work.

REFERENCES




- [1] D. Caballero, "Feature extraction algorithms from MRI to evaluate quality parameters on meat products by using data mining," *ELCVIA: electronic letters on computer vision and image analysis*, vol. 16, no. 2, pp. 1–4, 2018.
- [2] W.-T. Wu *et al.*, "Data mining in clinical big data: the frequently used databases, steps, and methodological models," *Military Medical Research*, vol. 8, no. 1, pp. 1–12, 2021, doi: 10.1186/s40779-021-00338-z.
- [3] S. Ghosh, A. Roy, and S. Chakraborty, "Support vector regression based metamodeling for seismic reliability analysis of structures," *Applied Mathematical Modelling*, vol. 64, pp. 584–602, 2018, doi: 10.1016/j.apm.2018.07.054.
- [4] R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, and K. I. Kroeker, "An overview of clinical decision support systems: benefits, risks, and strategies for success," *NPJ Digital Medicine*, vol. 3, no. 1, pp. 1–10, 2020, doi: 10.1038/s41746-020-0221-y.
- [5] K. Song, X. Zeng, Y. Zhang, J. D. Jonckheere, X. Yuan, and L. Koehl, "An interpretable knowledge-based decision support system and its applications in pregnancy diagnosis," *Knowledge-Based Systems*, vol. 221, pp. 1–12, 2021, doi: 10.1016/j.knsys.2021.106835.
- [6] D. Dave, H. Naik, S. Singhal, and P. Patel, "Explainable AI meets healthcare: a study on heart disease dataset," *Computer Science-Machine Learning*, pp. 1–23, 2020, doi: 10.48550/arXiv.2011.03195.
- [7] M. Tarawneh and O. Embarak, "Hybrid approach for heart disease prediction using data mining techniques," in *Advances in Internet, Data and Web Technologies*, Cham: Springer, 2019, pp. 447–454, doi: 10.1007/978-3-030-12839-5_41.
- [8] M. Aghamohammadi, M. Madan, J. K. Hong, and I. Watson, "Predicting heart attack through explainable artificial intelligence," in *Computational Science – ICCS 2019*, Cham: Springer, 2019, pp. 633–645, doi: 10.1007/978-3-030-22741-8_45.
- [9] S. Navdeep and J. Sonika, "Heart disease prediction system using hybrid technique of data mining algorithms," *International*

- Journal of Advance Research, Ideas and Innovations in Technology*, vol. 4, no. 2, pp. 982–987, 2018.
- [10] P. Singh, S. Singh, and G. S. P. -Jain, "Effective heart disease prediction system using data mining techniques," *International Journal of Nanomedicine*, vol. 13, pp. 121–124, 2018, doi: 10.2147/IJN.S124998.
 - [11] M. Sudha, "Evolutionary and neural computing based decision support system for disease diagnosis from clinical data sets in medical practice," *Journal of Medical Systems*, vol. 41, no. 178, pp. 1–10, 2017, doi: 10.1007/s10916-017-0823-3.
 - [12] A. Solanki and M. P. Barot, "Study of heart disease diagnosis by comparing various classification algorithms," *International Journal of Engineering and Advanced Technology*, vol. 8, no. 2, pp. 40–42, 2019.
 - [13] V. S. K. Reddy, P. Meghana, N. V. S. Reddy, and B. A. Rao, "Prediction on cardiovascular disease using decision tree and naïve bayes classifiers," *Journal of Physics: Conference Series*, vol. 2161, no. 1, pp. 1–7, 2022, doi: 10.1088/1742-6596/2161/1/012015.
 - [14] M. Benllarch, S. E. Hadaj, and M. Benhaddi, "Improve extremely fast decision tree performance through training dataset size for early prediction of heart diseases," in *2019 International Conference on Systems of Collaboration Big Data, Internet of Things & Security (SysCoBloTS)*, 2019, pp. 1–5, doi: 10.1109/SysCoBloTS48768.2019.9028026.
 - [15] N. Louridi, S. Douzi, and B. E. Ouahidi, "Machine learning-based identification of patients with a cardiovascular defect," *Journal of Big Data*, vol. 8, no. 1, p. 133, 2021, doi: 10.1186/s40537-021-00524-9.
 - [16] A. N. Sumin, "A new diagnostic algorithm for examining patients with suspected chronic coronary syndrome: questions remain?," *Rational Pharmacotherapy in Cardiology*, vol. 16, no. 3, pp. 474–480, 2020, doi: 10.20996/1819-6446-2020-06-14.
 - [17] K. Bahani, M. Moujabbar, and M. Ramdani, "An accurate fuzzy rule-based classification systems for heart disease diagnosis," *Scientific African*, vol. 14, p. 1019, 2021, doi: 10.1016/j.sciaf.2021.e01019.
 - [18] J. Kaur and B. S. Khehra, "Fuzzy logic and hybrid based approaches for the risk of heart disease detection: state-of-the-art review," *Journal of The Institution of Engineers (India): Series B*, vol. 103, no. 2, pp. 681–697, 2022, doi: 10.1007/s40031-021-00644-z.
 - [19] J. Martinsson, A. Schliep, B. Eliasson, and O. Mogren, "Blood glucose prediction with variance estimation using recurrent neural networks," *Journal of Healthcare Informatics Research*, vol. 4, no. 1, pp. 1–18, 2020, doi: 10.1007/s41666-019-00059-y.
 - [20] H. A. Salah and A. S. Ahmed, "Coronavirus disease diagnosis, care and prevention (COVID-19) based on decision support system," *Baghdad Science Journal*, vol. 18, no. 3, pp. 593–613, 2021, doi: 10.21123/bsj.2021.18.3.0593.
 - [21] A. S. Ahmed, H. A. Salah, and J. Q. Jameel, "Software development for first aid decision support system," *Iraqi Journal of Science*, vol. 61, no. 1, pp. 202–214, 2020, doi: 10.24996/ij.s.2020.61.1.23.
 - [22] R. J. Volk *et al.*, "Effect of a patient decision aid on lung cancer screening decision-making by persons who smoke," *JAMA Network Open*, vol. 3, no. 1, pp. 1–12, 2020, doi: 10.1001/jamanetworkopen.2019.20362.
 - [23] L. Moja *et al.*, "Effectiveness of a hospital-based computerized decision support system on clinician recommendations and patient outcomes," *JAMA Network Open*, vol. 2, no. 12, pp. 1–16, 2019, doi: 10.1001/jamanetworkopen.2019.17094.
 - [24] T. Goto, C. A. Camargo, M. K. Faridi, R. J. Freishtat, and K. Hasegawa, "Machine learning-based prediction of clinical outcomes for children during emergency department triage," *JAMA Network Open*, vol. 2, no. 1, pp. 1–14, 2019, doi: 10.1001/jamanetworkopen.2018.6937.
 - [25] R. Martinek *et al.*, "Advanced bioelectrical signal processing methods: past, present and future approach—part i: cardiac signals," *Sensors*, vol. 21, no. 15, pp. 1–32, 2021, doi: 10.3390/s211515186.
 - [26] N. Aljohani, F. Nadeem, M. Abumelha, and A. Hashbal, "Development of infection control surveillance system for intensive care unit: data requirements and guidelines," *International Journal of Intelligent Systems and Applications*, vol. 8, no. 6, pp. 19–26, 2016, doi: 10.5815/ijisa.2016.06.03.
 - [27] I. K. Mujawar and B. T. Jadhav, "Web-based fuzzy expert system for diabetes diagnosis," *International Journal of Computer Sciences and Engineering*, vol. 7, no. 2, pp. 995–1000, 2019, doi: 10.26438/ijcse/v7i2.9951000.
 - [28] I. Cricelli, E. Marconi, and F. Lapi, "Clinical decision support system (CDSS) in primary care: from pragmatic use to the best approach to assess their benefit/risk profile in clinical practice," *Current Medical Research and Opinion*, vol. 38, no. 5, pp. 827–829, 2022, doi: 10.1080/03007995.2022.2052513.
 - [29] H. Zha *et al.*, "Acceptance of clinical decision support system to prevent venous thromboembolism among nurses: an extension of the UTAUT model," *BMC Medical Informatics and Decision Making*, vol. 22, no. 1, pp. 1–12, 2022, doi: 10.1186/s12911-022-01958-8.
 - [30] A. M. Antoniad *et al.*, "Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review," *Applied Sciences*, vol. 11, no. 11, pp. 1–23, 2021, doi: 10.3390/app11115088.
 - [31] I. H. Sarker, "Machine learning: algorithms, real-world applications and research directions," *SN Computer Science*, vol. 2, no. 3, pp. 1–21, 2021, doi: 10.1007/s42979-021-00592-x.
 - [32] N. Kunhare, R. Tiwari, and J. Dhar, "Particle swarm optimization and feature selection for intrusion detection system," *Sādhanā*, vol. 45, no. 1, pp. 1–14, 2020, doi: 10.1007/s12046-020-1308-5.
 - [33] A. Z. Adamov, "Analysis of feature selection techniques for classification problems," in *2021 IEEE 15th International Conference on Application of Information and Communication Technologies (AICT)*, 2021, pp. 1–6, doi: 10.1109/AICT52784.2021.9620226.
 - [34] M. Alloghani, D. A. -Jumeily, J. Mustafina, A. Hussain, and A. J. Aljaaf, "A systematic review on supervised and unsupervised machine learning algorithms for data science," in *Supervised and Unsupervised Learning for Data Science*, Cham: Springer, 2020, pp. 3–21, doi: 10.1007/978-3-030-22475-2_1.
 - [35] Y. Komaru, T. Yoshida, Y. Hamasaki, M. Nangaku, and K. Doi, "Hierarchical clustering analysis for predicting 1-year mortality after starting hemodialysis," *Kidney International Reports*, vol. 5, no. 8, pp. 1188–1195, 2020, doi: 10.1016/j.ekir.2020.05.007.
 - [36] K. S. D. S. C. R. and B. P., "Association of abnormal liver function tests to outcomes in patients with a new diagnosis of heart failure with reduced ejection fraction in the outpatient clinic," *Journal of Cardiovascular Diseases & Diagnosis*, vol. 5, no. 6, pp. 1–4, 2017, doi: 10.4172/2329-9517.1000300.
 - [37] M. D. Sans, S. J. Crozier, N. L. Vogel, L. G. D'Alecy, and J. A. Williams, "Dietary protein and amino acid deficiency inhibit pancreatic digestive enzyme mRNA translation by multiple mechanisms," *Cellular and Molecular Gastroenterology and Hepatology*, vol. 11, no. 1, pp. 99–115, 2021, doi: 10.1016/j.jcmgh.2020.07.008.
 - [38] S. H. Zolfani and A. Derakhti, "Synergies of text mining and multiple attribute decision making: a criteria selection and weighting system in a prospective MADM outline," *Symmetry*, vol. 12, no. 5, pp. 1–18, 2020, doi: 10.3390/sym12050868.
 - [39] I. Tougui, A. Jilbab, and J. E. Mhamdi, "Heart disease classification using data mining tools and machine learning techniques," *Health and Technology*, vol. 10, no. 5, pp. 1137–1144, 2020, doi: 10.1007/s12553-020-00438-1.
 - [40] K. Das and R. N. Behera, "A survey on machine learning: concept, algorithms and applications," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 5, no. 2, pp. 1301–1309, 2017.




- [41] A. Bartschat, M. Reischl, and R. Mikut, "Data mining tools," *WIREs Data Mining and Knowledge Discovery*, vol. 9, no. 4, pp. 1–14, 2019, doi: 10.1002/widm.1309.
- [42] R. J. Roiger, *Data mining*. New York: Chapman and Hall/CRC, 2017, doi: 10.1201/9781315382586.
- [43] A. Sharma and B. Kaur, "A research review on comparative analysis of data mining tools, techniques and parameters," *International Journal of Advanced Research in Computer Science*, vol. 8, no. 7, pp. 523–529, 2017, doi: 10.26483/ijarcs.v8i7.4255.
- [44] J. Padarian, B. Minasny, and A. B. McBratney, "Machine learning and soil sciences: a review aided by machine learning tools," *SOIL*, vol. 6, no. 1, pp. 35–52, 2020, doi: 10.5194/soil-6-35-2020.
- [45] A. A. Abdulrahman and M. K. Ibrahim, "Evaluation of DDoS attacks detection in a new intrusion dataset based on classification algorithms," *Iraqi Journal of Information and Communications Technology*, vol. 1, no. 3, pp. 49–55, 2018, doi: 10.31987/ijict.1.3.40.
- [46] N. Mishra, R. K. Singh, and S. K. Yadav, "Detection of DDoS vulnerability in cloud computing using the perplexed bayes classifier," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–13, 2022, doi: 10.1155/2022/9151847.
- [47] A. K. Pandey, D. S. Rajpoot, and D. S. Rajpoot, "A comparative study of classification techniques by utilizing WEKA," in *2016 International Conference on Signal Processing and Communication (ICSC)*, 2016, pp. 219–224, doi: 10.1109/ICSPCom.2016.7980579.
- [48] I. Sharafaldin, A. H. Lashkari, S. Hakak, and A. A. Ghorbani, "Developing realistic distributed denial of service (DDoS) attack dataset and taxonomy," in *2019 International Carnahan Conference on Security Technology (ICCST)*, 2019, pp. 1–8, doi: 10.1109/CCST.2019.8888419.
- [49] A. Sudugala, W. Chanuka, A. M. Eshan, U. C. Bandara, and K. Abeywardena, "WANHEDA: a machine learning based DDoS detection system," in *2020 2nd International Conference on Advancements in Computing (ICAC)*, 2020, pp. 380–385, doi: 10.1109/ICAC51239.2020.9357130.
- [50] H. Sifaou, A. Kammoun, and M.-S. Alouini, "A precise performance analysis of support vector regression," in *Proceedings of the 38th International Conference on Machine Learning*, 2021, vol. 139, pp. 9671–9680.

BIOGRAPHIES OF AUTHORS



Ahmed Shihab Ahmed    is a computer scientist specialized in the field of image processing and decision support systems. He received the four-year B.Sc. degree in Computer Science in 2000 from Al-Rafidain University College, Iraq. In 2015, he concluded a Master in Computer Science (MCS) from Middle East University, Jordan. He has been working as a programmer at University of Baghdad from 2004 until 2014 and then worked as a lecturer at University of Baghdad from 2015 until now. His main research interests include: artificial neural network, image processing, and decision support systems. He can be contacted at email: ahmedshihabinfo@conursing.uobaghdad.edu.iq.



Hussein Ali Salah    received the four-year B.Sc. degree in Computer Science in 2000 from Al- Rafidain University College, Iraq. In 2004, he concluded a Master in Computer Science (MCS) from Baghdad University, college of science. He received the Ph.D. degree in Computer Science IT in 2016 from Politehnica' University of Bucharest, Bucharest, Romania. His main research interests include data mining, decision support system, web design and intelligent DSS. He has worked as an assistant Prof. in the Department of Computer Systems, Middle Technical University, Technical Institute-Suwaira, Wasit, Iraq from 2016 until now. He can be contacted at email: hussein_tech@mtu.edu.iq.